COMPUTER SCIENCE: Where Are the Exemplars?
Marc Mézard, *et al.*
Science **315**, 949 (2007);
DOI: 10.1126/science.1139678

(*14*). Future work will surely focus on why more of apparently the same neurons seem to have a different function.

References

1. J. O'Keefe, *Exp. Neurol.* **51**, 78 (1976).
2. M. A. Wilson, B. L. McNaughton, *Science* **261**, 1055 (1993).
3. J. K. Leutgeb, S. Leutgeb, M.-B. Moser, E. I. Moser, *Science* **315**, 961 (2007).
4. T. J. Wills, C. Lever, F. Cacucci, N. Burgess, J. O'Keefe, *Science* **308**, 873 (2005).
5. J. J. Hopfield, *Proc. Natl. Acad. Sci. U.S.A.* **79**, 2554 (1982).
6. J. K. Leutgeb *et al.*, *Neuron* **48**, 345 (2005).
7. R. M. Hayman, S. Chakraborty, M. I. Anderson, K. J. Jeffery, *Eur. J. Neurosci.* **18**, 2825 (2003).
8. K. D. Harris, J. Csicsvari, H. Hirase, G. Dragoi, G. Buzsaki, *Nature* **424**, 552 (2003).
9. E. Pastalkova *et al.*, *Science* **313**, 1141 (2006).
10. J. R. Whitlock, A. J. Heynen, M. G. Shuler, M. F. Bear, *Science* **313**, 1093 (2006).
11. R. U. Muller, J. L. Kubie, *J. Neurosci.* **7**, 1951 (1987).
12. M. H. Fyhn, T. F. Hafting, A. Treves, E. I. Moser, M. B. Moser, *Soc. Neurosci. Abstr.* **68**, 9 (2006).
13. M. K. Chawla *et al.*, *Hippocampus* **15**, 579 (2005).
14. E. Gould, A. Beylin, P. Tanapat, A. Reeves, T. J. Shors, *Nat. Neurosci.* **2**, 260 (1999).

110.1126/science.1139146

COMPUTER SCIENCE

Where Are the Exemplars?

A fast way of finding representative examples in complex data sets may be applicable to a wide range of difficult problems.

Marc Mézard

As a flood of data pours from scientific and medical experiments, researchers crave more efficient computational methods to organize and analyze it. When dealing with large, noisy data sets, scientists often use a computational method that looks for data clusters. In the case of gene expression with tens of thousands of sequences, for example, the clusters would be groups of genes with similar patterns of expression. On page 972 of this issue, Frey and Dueck propose a new method for finding an optimal set of clusters (*1*). Their algorithm detects special data points called exemplars, and connects every data point to the exemplar that best represents it. In principle, finding an optimal set of exemplars is a hard problem, but this algorithm is able to efficiently and quickly handle very large problems (such as grouping 75,000 DNA segments into 2000 clusters). An analysis that would normally take hundreds of hours of computer time might now be done in a few minutes.

Detecting exemplars goes beyond simple clustering, as the exemplars themselves store compressed information. An example with a broad range of possible applications is found in the statistical analysis of language. For instance, take your last scientific paper (and no, I don't really suggest that it is a large, noisy data set) and consider each sentence to be a data point. The similarity between any two sentences can be computed with standard information theory methods (that is, the similarity increases when the sentences include the same words). Knowing the similarities, one can detect the exemplary sentences in the paper, which provide an optimally condensed description. If you are a hasty reader, you can thus go directly to Fig. 4 of Frey and Dueck's report and find the best summary of their own paper in four sentences. But understanding the method requires a bit more effort.

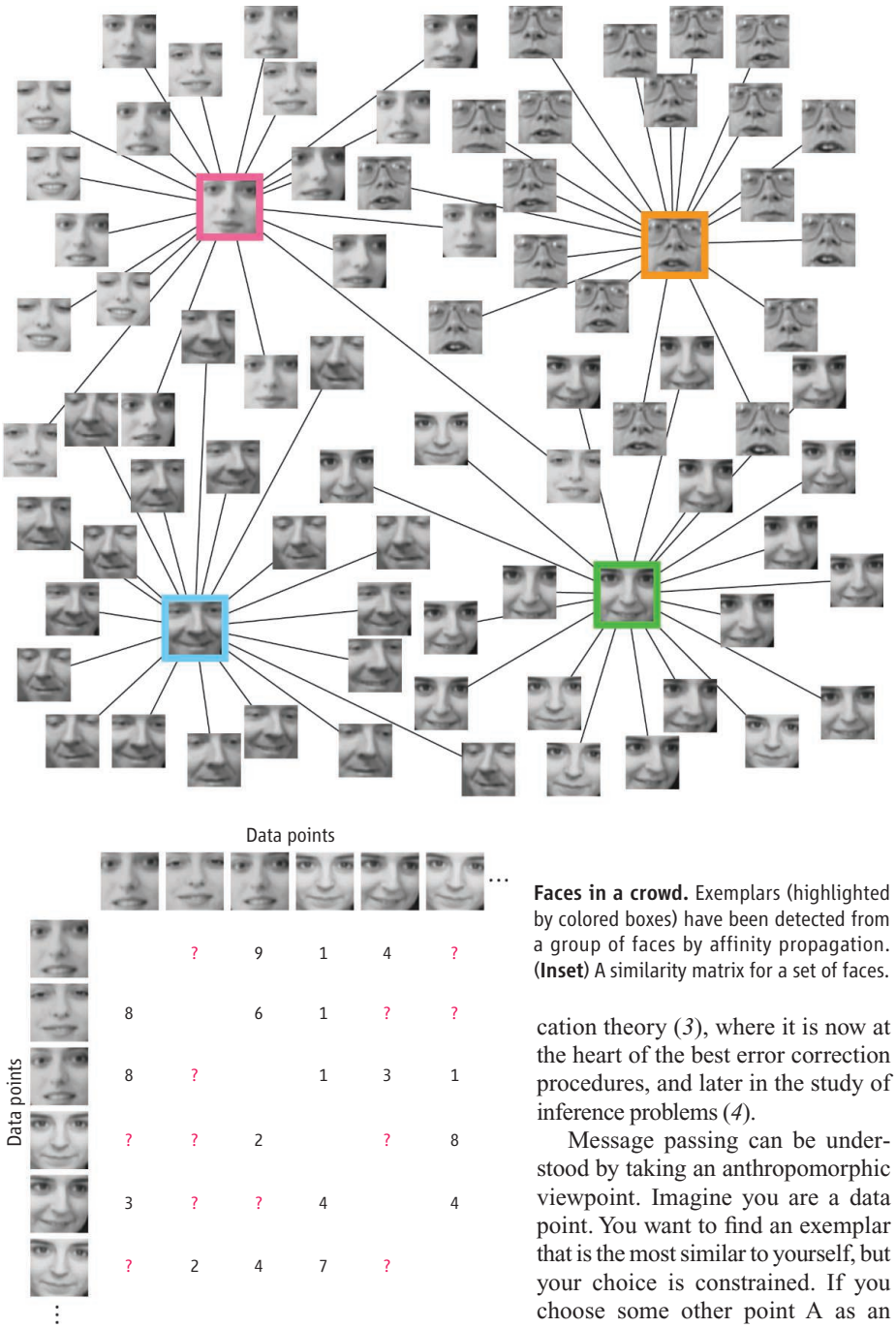
Such methods start with the construction of a similarity matrix, a table of numbers that establishes the relationship of each data point to every other data point. As we saw in the semantics example, $S(B, A)$ is a number that measures how well the data point A represents point B [and it is not necessarily equal to $S(A, B)$]. The optimal set of exemplars is the one for which the sum of similarities of each point to its exemplar is maximized. In the usual clustering methods (*2*), one decides a priori on the number of exemplars, and then tries to find them by iterative refinement, starting from a random initial choice.

The method of Frey and Dueck, called affinity propagation, does not fix the number of exemplars. Instead, one must choose for each point B a number $P(B)$ that characterizes

CREDIT: SCALA/ART RESOURCE

Caravaggio's "Vocazione di San Matteo." How to choose an exemplar through message passing. The messages are exchanged in the directions of the fingers and of the glances, leading to the recognition of San Matteo as the "exemplar."

The author is at the Centre National de la Recherche Scientifique and Laboratoire de Physique Théorique et Modeles Statistiques, Université Paris Sud, 91405 Orsay, France. E-mail: mezard@lptms.u-psud.fr



Faces in a crowd. Exemplars (highlighted by colored boxes) have been detected from a group of faces by affinity propagation. (**Inset**) A similarity matrix for a set of faces.

the a priori knowledge of how good point B is as an exemplar. In most cases all points are equally suitable, so all the numbers take the same value P. This quantity provides a control parameter: The larger P, the more exemplars one is likely to find.

Affinity propagation is known in computer science as a message-passing algorithm (see the first figure) and it aims at maximizing the net similarity. It is in fact an application of a method called "belief propagation," which was invented at least twice: first in communication theory (*3*), where it is now at the heart of the best error correction procedures, and later in the study of inference problems (*4*).

Message passing can be understood by taking an anthropomorphic viewpoint. Imagine you are a data point. You want to find an exemplar that is the most similar to yourself, but your choice is constrained. If you choose some other point A as an exemplar, then A must also decide to be its own exemplar. This creates one constraint per data point, establishing a large network of constraints that must all be satisfied. When the net similarity is maximized with all constraints satisfied, the set of actual exemplars emerges.

Now imagine that next to each point stands a guardian angel telling whether someone else has chosen that point as an exemplar or not. An approximate solution of the complicated web of conflicting constraints is obtained by having all of these characters talk to each other. At a given time, all angels send messages to all points, and all points answer to all angels. One data point tells the angel of every other point his ranked list of favorite exemplars. An angel tells every other point the degree of compatibility of his list with the angel's constraints. Every sent message is evaluated through a simple computation on the basis of the received messages and the similarity matrix. After several message-passing rounds, all the characters reach an agreement and every point knows its exemplar. In practice, the running time of this algorithm scales linearly with the number of similarities.

As an example, affinity propagation can be a powerful method to extract representative faces from a gallery of images (see the second figure). The input is a list of numerical similarities between pairs of data points, which may be measured, computed using a model, or, in the present example, set by visual inspection (missing similarity values indicated with question marks are accepted by the algorithm). Each face is a data point that exchanges messages with all other faces and their guardian angels. After a few iterations of message passing, a global agreement is reached on the set of exemplars.

Such message-passing methods have been shown to be remarkably efficient in many hard problems that include error correction, learning in neural networks, computer vision, and determining the satisfiability of logical formulas. In many cases they are the best available algorithms, and this new application to cluster analysis looks very powerful. Understanding their limits is a main open challenge. At the lowest level this means controlling the convergence properties or the quality of the approximate solutions that they find. A more ambitious goal is to characterize the problems where they can be useful. The concepts and methods developed in statistical physics to study collective behavior offer the most promising perspective in this respect. In physics terms, belief propagation (and therefore affinity propagation) is a mean field–type method (*5*). That is, the complex interaction of a given object (a data point) with all of the others is approximated by an average effective interaction. Although this works well in most cases, it may get into trouble when the system gets close to a phase transition (*6*), where some correlations become extremely long-ranged. The appropriate modification, which requires using more sophisticated messages, has been worked out in some special cases (*7*), but again its full range of applicability is still to be found.

Along with its pedagogical virtue, the anthropomorphic explanation of message passing also underlines its main features. This

strategy can find an excellent approximate solution to some of the most difficult computational problems with a very simple recipe: It uses basic messages which are exchanged in a distributed system, together with simple update rules that are purely local. This realizes in practice a new scheme of computation, based on distributed simple elements that operate in parallel, in the spirit of neurocomputation. One might expect to find that some of its principles are at work in living organisms or social systems. Each new successful application of message passing, such as affinity propagation, thus adds to our understanding of complex systems.

References

1. B. J. Frey, D. Dueck, *Science* **315**, 972 (2007); published online 11 January 2007 (10.1126/science.1136800).
2. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. Le Cam, J. Neyman, Eds. (Univ. of California Press, Berkeley, CA, 1967), p. 281.
3. R. G. Gallager, *Low Density Parity Check Codes* (MIT Press, Cambridge, MA, 1963).
4. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Morgan Kaufmann, San Mateo, CA, 1988).
5. J. S. Yedidia, W. T. Freeman, Y. Weiss, *IEEE Trans. Infor. Theory* **51**, 2282 (2005).
6. O. Dubois, R. Monasson, B. Selman, R. Zecchina, Eds., special issue on Phase Transitions in Combinatorial Problems, *Theor. Comp. Sci.* **265** (2001).
7. M. Mézard, G. Parisi, R. Zecchina, *Science* **297**, 812 (2002); published online 27 June 2002 (10.1126/science.1073287).

10.1126/science.1139678

GEOLOGY

On the Origins of Granites

Granites make up a large part of the continental crust. New data reveal their complex and diverse formation history, calling for a revision of the geological histories of many granites.

John M. Eiler

Geology spent the 19th and much of the 20th century fighting a scientific civil war over the origin of granites—the coarsely crystalline, feldspar-rich rocks that make such excellent building stones and kitchen counters. The ultimate losers (*1*) held that granites precipitated from aqueous fluids that percolate through the crust, or formed by reaction of preexisting rocks with such fluids; the winners (*2*) recognized that granites crystallized from silicate melts.

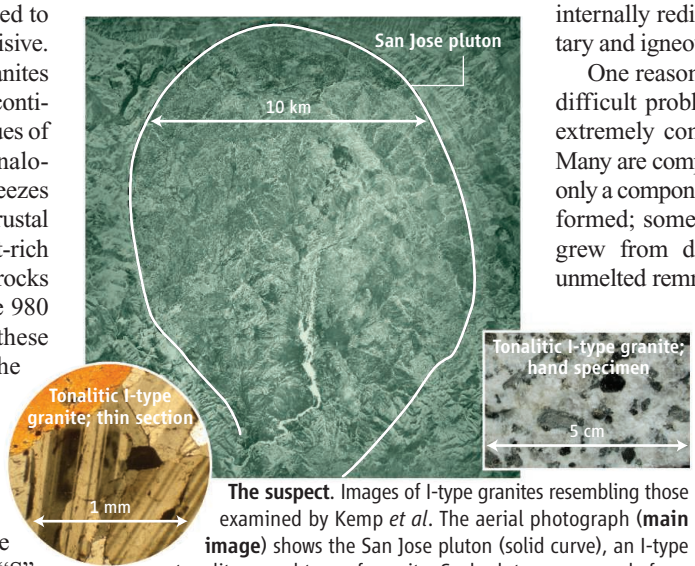
Yet, the resolution of this argument led to various others that remain almost as divisive. Are the silicate melts that give rise to granites partial melts of preexisting rocks in the continental crust, or are they instead the residues of crystallizing mantle-derived basalts, analogous to the brine that is left when ice freezes out of salty water? If granites form by crustal melting, do they come from the sediment-rich upper crust or from preexisting igneous rocks that dominate the lower crust? On page 980 of this issue, Kemp *et al*. (*3*) examine these questions through the lens of two of the newest analytical tools developed for the earth sciences.

Clear answers to the above questions have been found previously for some extreme types of granite. There is little debate that upper-crustal sediments are the sources of S-type granites (*4*) (where "S" stands for sediment) and that mantle-derived basalts give rise to M-type granites (*5*) ("M" for mantle). However, members of a third class—the I-type (*4*)—are abundant, widely distributed, and diverse, and their origins are up for grabs. A popular view holds that these granites are melts of deep-crustal igneous rocks (hence the "I" for igneous) (*4*, *6*). A minority dissenting view suggests that they are instead largely mantle-derived and only modified by passage through the crust (*7*).

The stakes in this argument are high: I-type granites (or their metamorphosed or eroded derivatives) make up a large fraction of the continental crust. Therefore, our thoughts regarding their origins are key to understanding the mechanisms by which the continents differentiate from the rest of the silicate earth, and the consequences of that differentiation for the composition of the mantle. If I-type granites are descended from basalts, then their formation represents net growth of the continents and net removal from the mantle of elements that are highly concentrated in the crust (such as the heat-producing radioactive isotopes, ^{40}K and ^{238}U). If, instead, they form by melting preexisting crustal rocks, they represent a mechanism by which the continents internally redistribute their various sedimentary and igneous constituents.

One reason the origin of granite is such a difficult problem is that these rocks can be extremely complicated (see the figure) (*8*). Many are composed of minerals that represent only a component of the melts from which they formed; some are mixtures of minerals that grew from different melts; some contain unmelted remnants of their sources; and individual minerals often have heterogeneous chemical and isotopic compositions, reflecting the evolution of their parental magmas over the course of their crystallization.

Kemp *et al*. (*3*) examine the origin and evolution of I-type granites from the Lachlan belt in Australia. Their work draws on several recent microanalytical innovations, including high-precision, in situ measurements of oxygen isotope ratios with a large-radius ion microprobe and in situ measurements of hafnium isotopes using laser ablation joined with an inductively coupled plasma mass spec-



The suspect. Images of I-type granites resembling those examined by Kemp *et al*. The aerial photograph (**main image**) shows the San Jose pluton (solid curve), an I-type tonalite, or subtype of granite. Such plutons commonly form kilometer-scale bodies intruded into rocks of the upper crust. Kemp *et al*. suggest that assimilation of enveloping rocks influences the compositions of such bodies. The insets show a specimen of a similar tonalite from the Chihuahua Valley, California. The visible light photograph (**right inset**) reveals dark laths of amphibole and hexagonal crystals of biotite embedded in a white matrix of interlocking feldspar and quartz. The transmitted-light photomicrograph (**left inset**) shows twinning, compositional zoning, overgrowths, and inclusions in plagioclase (complex light and dark pattern), adjacent to a crystal of amphibole (brown). The micro-analytical techniques employed by Kemp *et al*. aim to avoid artifacts that arise from mixing different components of these compositionally and texturally complex rocks.

CREDIT: MAIN IMAGE, J. D. MURRAY AND L. T. SILVER; INSETS, J. M. EILER

The author is in the Division of Geological and Planetary Sciences, California Insititute of Technology, Pasadena, CA 91125, USA. E-mail: eiler@gps.caltech.edu